

Comparative sequence analysis of imprinted genes between human and mouse to reveal imprinting signatures[☆]

Zhining Wang,¹ Hongtao Fan,¹ Howard H. Yang,¹ Ying Hu,
Kenneth H. Buetow, and Maxwell P. Lee^{*}

Laboratory of Population Genetics, National Cancer Institute, Bethesda, MD 20892, USA

Received 7 May 2003; accepted 4 September 2003

Abstract

We performed a comparative genomic sequence analysis between human and mouse for 24 imprinted genes on human chromosomes 1, 6, 7, 11, 13, 14, 15, 18, 19, and 20. The MEME program was used to search for motifs within conserved sequences among the imprinted genes and we then used the MAST program to analyze for the presence or absence of motifs in the imprinted genes and 128 nonimprinted genes. Our analysis identified 15 motifs that were significantly enriched in the imprinted genes. We generated a logistic regression model by combining multiple motifs as input variables and the 24 imprinted genes and the 128 nonimprinted genes as a training set. The accuracy, sensitivity, and specificity of our model were 98, 92, and 99%, respectively. The model was further validated by an open test on 12 additional imprinted genes. The motifs identified in this study are novel imprinting signatures, which should improve our understanding of genomic imprinting and the role of genomic imprinting in human diseases.

Published by Elsevier Inc.

Keywords: Imprinting; Motif; Comparative genomics; Logistic regression model

Introduction

Genomic imprinting is an unusual mechanism of gene regulation that results in preferential expression of one specific parental allele of a gene. Abnormal imprinting can cause human diseases such as Beckwith–Wiedemann syndrome, Prader–Willi syndrome, or Angelman syndrome [1–3]. Loss of imprinting is often associated with human cancers [4,5]. Although the exact mechanism of genomic imprinting is still largely unknown, differentially methylated CpG islands, imprinted antisense transcripts, and insulators may play important roles in the regulation of imprinting [6–8]. Most of the imprinted genes are located in the imprinting domains [9]. However, some genes in the imprinting domain can escape imprinting regulation [10]. Many imprinted

genes are scattered throughout the human genome. Therefore, it is likely that local *cis*-elements as well as chromatin structure control genomic imprinting.

Since patterns of gene regulation and the corresponding regulatory elements are often conserved across species, sequence comparison between human and mouse is a powerful approach to identify regulatory sequences [11]. Such comparative sequence analysis has already identified a number of conserved sequences and novel imprinted genes in human 11p15 [12] and the Dlk1–Gtl2 locus [13,14]. In this report, we extend the comparative genomic sequence analysis to the known imprinted genes in the entire human genome. We then go on to identify motifs shared among the conserved sequences and discover a new imprinting signature.

Results

Conserved sequences between human and mouse imprinted genes

We set out to identify novel sequence motifs that are associated with imprinted genes. Our computation method is

[☆] Supplementary data for this article may be found on ScienceDirect, at doi: 10.1016/j.ygeno.2003.09.007.

^{*} Corresponding author. Laboratory of Population Genetics, National Cancer Institute, 41 Library Drive, Room D702C, Bethesda, MD 20892, USA. Fax: +1-301-402-9325.

E-mail address: leemax@mail.nih.gov (M.P. Lee).

¹ These authors contributed equally to this study.

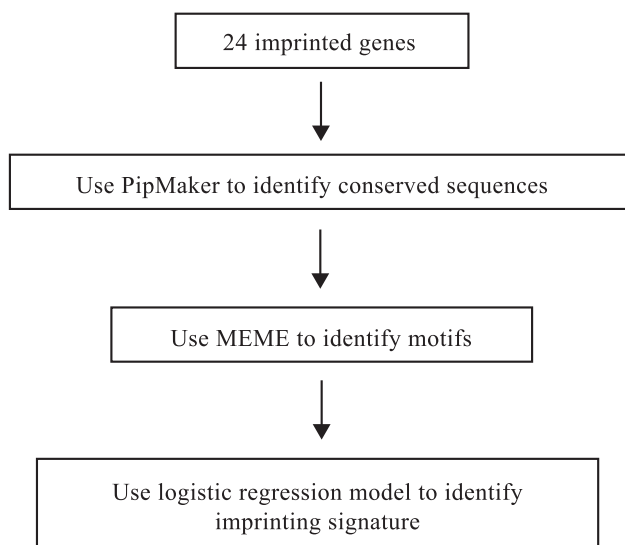


Fig. 1. A schema for computational analysis of imprinting signature. We started with 41 imprinted genes from human. We were able to find both human and mouse genomic sequences for 36 of 41 imprinted genes. Twenty-four imprinted genes were analyzed through PipMaker and MEME programs. The MAST program was performed for the 36 imprinted genes as well as for 128 nonimprinted genes. The 24 imprinted genes and the 128 nonimprinted genes were used as a training set, whereas the remaining 12 imprinted genes were used as a testing set.

depicted in Fig. 1. Regulatory elements tend to locate on the conserved sequences [11]. Therefore, we searched conserved sequences between human and mouse imprinted genes using the PipMaker program [15]. We started with a list of 41 known imprinted genes that we generated from literatures (Supplemental Table 1). Genomic sequences of these 41 imprinted genes (including their 10-kb upstream and 10-kb downstream sequences) were retrieved from <ftp://ftp.ncbi.nih.gov/genomes/>. We were able to find both human and mouse sequences for 36 imprinted genes, 24 of which were used as a training set and 12 of which were used as a testing set. The PipMaker program was used to align human and mouse genomic DNA sequences. An example of

extensive sequence homology in the upstream 10-kb region of NNAT is shown in Fig. 2.

Motifs associated with imprinted genes

It is of great interest to know whether imprinted genes share common motifs. We used the MEME program [16] to search motifs in the conserved noncoding sequences among the human imprinted genes. This analysis identified 16 motifs (Table 1). Motifs 1–4 are located in the upstream regions of the imprinted genes while motifs 5–8 and motifs 9–16 are located in the downstream and intron regions of the imprinted genes, respectively. The lengths of these motifs ranged from 19 to 50 bp (Table 1). A motif defined by the MEME program is not just a consensus sequence, but a position-specific probability matrix, which has probabilities associated with each base at each position [16]. We then used the MAST program [17] to search for the presence of these motifs in the 24 imprinted genes as well as in 128 nonimprinted genes, which were identified in our previous study (Supplemental Table 2) [18]. Motifs 1–4, motifs 5–8, and motifs 9–16 were searched for in the upstream, downstream, and intron sequences of the 24 imprinted genes and the 128 nonimprinted genes. Fifteen of the 16 motifs were found to be significantly associated with the 24 imprinted genes ($p < 0.05$, Fisher exact test, Table 1). Interestingly, motif 2 was present only in the imprinted genes on chromosome 11p15 and absent from any other chromosomes. Unlike motif 2, the other motifs found in this study were present on multiple chromosomes. For example, motif 12 was found in 15 imprinted genes on eight different chromosomes (Table 1).

Imprinting signatures

It has been suggested that imprinted genes share some common features [12,19]. In the present study, we found that all motifs except motif 5 were significantly enriched in the imprinted genes. Based on the distribution of the motifs

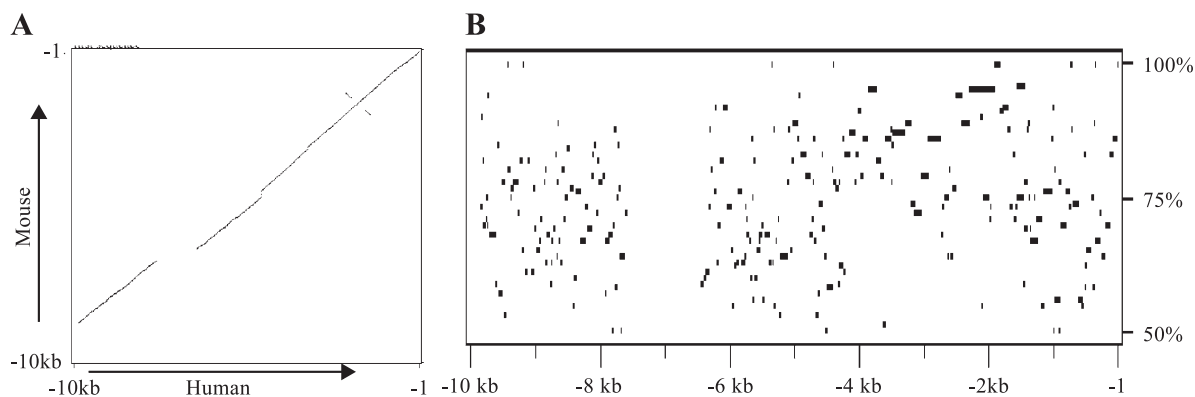


Fig. 2. Conserved sequences between human and mouse homologous NNAT gene at the upstream 10-kb region. (A) The dot-plot output from the PipMaker program. (B) The percentage identity plot output from the PipMaker program. The x axis indicates the position. The transcription initiation site is at the 0 position. The y axis indicates the percentage of identity between human and mouse.

Table 1
Summary of motifs found in the known imprinted genes in human

Motif	Width	Counts in 24 imprinted genes	Counts in 128 nonimprinted genes	<i>p</i> value	Consensus sequence
1	25	13	0	1.14×10^{-12}	GAGGGTGGGGGGCAGGCCGAGGAGG
2	43	3	0	0.003	AGCCACATCTCTGGGTATAGAGTGAATGTCCAGTTTTCATTA
3	26	12	6	1.49×10^{-7}	AGAATAAATGAAAAAAAAAATAAAAG
4	26	2	0	0.024	TGTCGTGGTGACAGCACTGTGCCCAT
5	19	2	5	0.304	TCCACCCACCCACTCACCC
6	26	4	0	4.97E-04	TGGAGGGGCGAGTCCGGCTCCTGGGGG
7	26	8	0	1.26E-07	ATATTATGTTTTTTTTCATTTTCAAT
8	48	6	0	8.68E-06	GTGGTGACGATGGTGAGGTAGAGGAAGAGGGGGGTGCACTCCACCGAG
9	35	3	0	0.003	TCTAGCCCTCCATCTTAGCTCTTGGGCTCCCCAGC
10	36	20	15	4.51E-12	GCCCCGCCCCGCCCCCTCCCGCGCCGCGGCCGC
11	36	17	126	3.10E-05	ATTTTTTATTTTATTTTATTTTTTTTTTAAAA
12	35	14	2	9.40E-12	CAAGCTGATGAAGAAGATGCTGAACAAGAAGAAGA
13	19	4	0	4.97E-04	GGCCTGCCCTCCATCTTAG
14	50	6	0	8.68E-06	CTGGAATCCACCGACGCCGCTACCTGCAAACCACCTTCGGGGTCTCCA
15	50	2	0	0.024	GACTGCGCTACAACCTTCGAGTGGCACACCGCGGCTGCCTGCCCGAAGGA
16	26	4	0	4.97E-04	GATGCAGAAATTGAAGACCCAGAAGA

One-tailed Fisher's exact test was used. The sequences denote the consensus sequences. The position-specific probability matrix was used for database search.

among the 24 imprinted genes and the 128 nonimprinted genes, we developed a logistic regression model that was able to distinguish imprinted genes from nonimprinted genes (see Methods). We generated the logistic regression model by combining 12 motifs as input variables. The selection of the 12 motifs is described in detail under Methods. The 24 imprinted genes and the 128 nonimprinted genes were used as a training set. The following model was selected by using Akaike's information criterion (AIC)

$$p = 1/(1 + \exp(7.1 - 4.8 \times M3 - 12.2 \times M7 - 4.2 \times M10 - 4.9 \times M12 - 12.1 \times M13 - 12 \times M16)).$$

Our model correctly assigned 127 of the 128 nonimprinted genes and 22 of the 24 imprinted genes in the training set. Let T_p , T_n , F_p , and F_n be the numbers of true positives, true negatives, false positives, and false negatives, respectively, determined by comparing the model prediction to the actual imprinted and nonimprinted labels. We defined three performance measurements: accuracy = $(T_p + T_n)/(T_p + T_n + F_p + F_n)$, sensitivity = $T_p/(T_p + F_n)$, and specificity = $T_n/(T_n + F_p)$. The accuracy, sensitivity, and specificity of the above model are 98, 92, and 99%, respectively. To validate the performance of the model, we randomly select two-thirds of the 24 imprinted and the 128 nonimprinted genes as a training set and the remaining one-third of the imprinted and the nonimprinted genes as a testing set. A model was constructed based on the training set and the performance of the model was evaluated for both training and testing sets. We repeated this procedure 100 times. The performance histograms in the training and testing sets are shown in Figs. 3A and 3B. Accuracy and specificity of validation on the testing set are over 86 and 95%, respectively. If we set the probability threshold >0.6 , 22 of 24 imprinted genes and 127 of 128 nonimprinted genes were

correctly assigned (Fig. 3C). To further validate the model, we performed an open test on an additional 12 imprinted genes, which were set aside as a testing set as described earlier. The model is able to assign high probability scores to 8 of the 12 imprinted genes.

Discussion

In this paper, we carried out a comparative sequence analysis to identify conserved sequences. We then went on to identify motifs that were shared among imprinted genes. These motifs were used in logistic regression analysis to discover the imprinting signature.

Comparative sequence analysis is a powerful way to identify regulatory elements [11]. It has been used to identify conserved sequences in human 11p15 [12] and the Dlk1–Gtl2 locus [13,14]. However, the comparative sequence analysis for all imprinted genes in the entire human and mouse genomes has not been performed. We have further analyzed the conserved nonexonic sequences for the presence of common motifs using the MEME program [16]. The output from MEME is a weight matrix, which is then used to search for the presence of the motif in the local database containing the known imprinted genes and nonimprinted genes using the MAST program. The frequency distribution of the motifs in both the imprinted gene set and the non-imprinted gene set can be used to assess the significance of the association of the motifs with imprinted genes.

We used supervised learning to discover the imprinting signature. The training set contained 24 known imprinted genes and 128 known nonimprinted genes. The 128 non-imprinted genes came from our recent experimental study using a HuSNP chip to analyze allele-specific gene expression [18]. These 128 genes were selected because both alleles were expressed nearly equally. We used a logistic

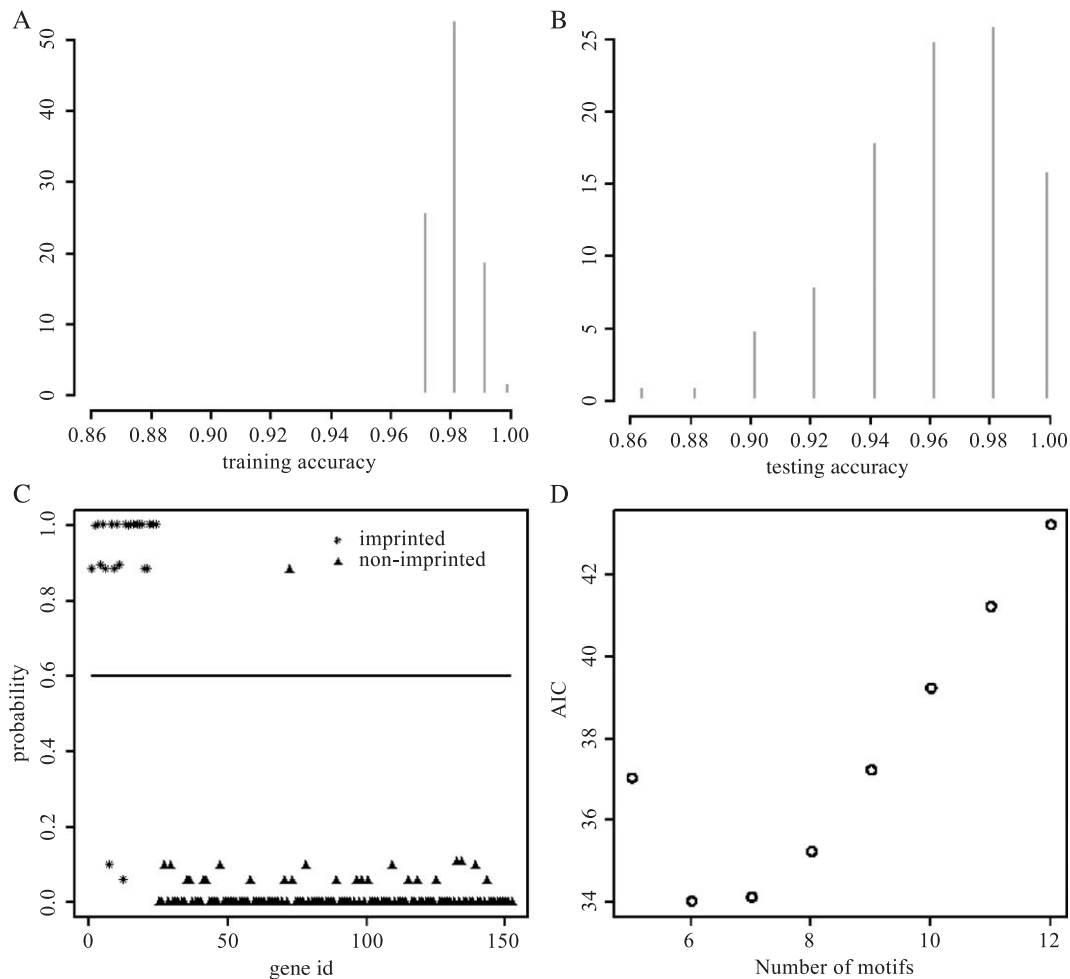


Fig. 3. Performance of the logistic regression model. (A and B) The histograms of training and testing accuracies in cross-validation. (C) The probability scores of the known imprinted and nonimprinted genes. (D) The AIC values of the models found in the stepwise searching process. The AIC value of the 6-motif set (3,7,10,12,13,16) is the global minimum among all possible subsets of the 12-motif set (1,2,3,6,7,8,9,10,12,13,14,16).

regression function to model the probability of a gene being imprinted based on its association with a set of motifs in the upstream, intronic, and downstream regions of this gene. We selected 6 optimal motifs from 12 motifs as predictor variables by minimizing AIC. The performance of the logistic regression model with the 6 motifs as inputs was evaluated by cross-validation using 2/3 of the data set as training and the rest of the data set as testing. The performance of the logistic regression model derived from 24 imprinted genes and 128 nonimprinted genes was further validated by the use of 12 imprinted genes that were different from those 24 used in the training. The model was able to assign high probability scores to 8 of 12 imprinted genes. This is significant considering that only 1% of the 128 nonimprinted genes has a high probability score and estimated imprinted genes represent about 1–2% of mammalian genes.

A number of genetic elements have been known to play important roles in genomic imprinting (for a recent review, see [20]). The imprinting control region (ICR) was initially identified in a study of Prader–Willi syndrome and Angel-

man syndrome as the region that was frequently deleted in the patients. It was subsequently found that imprinted genes are often associated with differentially methylated regions, which are usually located in CpG islands. The enhancer in the H19 gene downstream region was initially shown to be important for the reciprocal imprinting of IGF2 and H19. The current model of imprinting control in the IGF2 and H19 involves a silencer, which blocks the effect of the enhancer on IGF2. Differential methylation of the silencer, CpG islands, and promoter controls imprinting of IGF2 and H19. To understand the relationship between newly identified motifs and the known imprinting elements, we have mapped the relevant motifs along with the imprinting elements in IGF2, H19, KCNQ1, and SNRPN (Fig. 4). None of the motifs is mapped to the IGF2/H19 enhancer. In general, newly discovered motifs do not seem to associate with any known imprinting elements. Motif 3 appears twice in the 23-kb region encompassing upstream, gene, and downstream regions of H19. One overlaps with the imprinting control region and the other does not. There are 36 motifs in the 27-kb region of IGF2 from exon 3 to exon 9. Three motifs 13

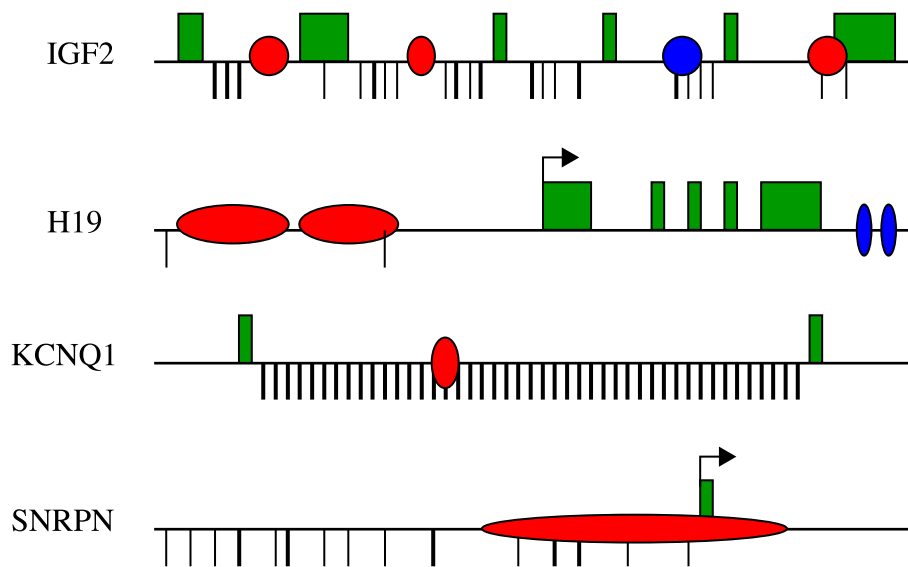


Fig. 4. Mapping of motifs and imprinting elements in IGF2, H19, KCNQ1, and SNRPN. Green rectangles represent exons, while circles and ellipses denote imprinting elements such as ICR, DMR, enhancer, silencer, and CpG islands. Vertical lines are for motifs, thin lines for single site and thick lines for multiple sites. Arrows show direction of transcription. Transcription is from left to right. IGF2—The 27-kb region from exon 3 to exon 9 of IGF2 is shown here. Three differentially methylated regions are marked with red circles. The DMR in intron 4 and intron 8/exon 9 are located within a CpG island. The silencer in intron 7 contains a conserved inverted repeat (blue circle). DMR in intron 4 also contains silencing activity and is located in the promoter region of IGF2-AS. The coordinates in NT_028310.10 for the four imprinting elements are 929,830..930,058 (DMR in intron 3), 921,066..922,725 (DMR in intron 4), 915,693..916,017 (insulator in intron 7), and 914,246..915,027 (DMR in intron 8 and exon 9). The vertical lines mark the positions of motif 13. H19—The 23-kb region encompassing upstream, gene, and downstream regions of H19 is shown here. Blue ellipses represent enhancers and red ellipses are for ICR that contain silencer, DMR, conserved repeats, and CTCF binding sites. The coordinates in NT_028310.10 for the four imprinting elements are 783,739..784,828 (the first DMR, which contains A2, B5, B6, and B7), 781,769..782,727 (the second DMR which contains A1, B1, B2, and B3), 771,584..771,760 (enhancer), and 769,877..770,181 (enhancer). The vertical lines mark the positions of motif 3. KCNQ1—A 107-kb region of KCNQ1 intron 10 is shown here. The coordinates in NT_028310.10 for ICR are 1,480,537..1,482,671 (DMR that contains CpG island). The vertical lines mark the positions of motif 13. SNRPN—The 10-kb upstream region of SNRPN is shown here. The coordinates in NT_026446.12 for ICR are 1,631,334..1,635,609. The vertical lines mark the positions of the motif 3.

are located in an insulator containing a conserved inverted repeat and one motif 13 is located in DMR2 near exon 9. The rest of the motifs 13 are scattered around. Motif 13 also occurs frequently in KCNQ1 intron 10, where an ICR was found. There are 127 motifs 13 in a 107-kb region of KCNQ1 intron 10 and 2 of them are located in the 2-kb CpG island containing ICR. Despite the high frequency of motif 13 in IGF2 and KCNQ1, it was absent in any of the 128 nonimprinted genes. Motif 3 occurs 24 times in the 10-kb upstream region of SNRPN and 8 of them are located in the ICR. Our newly identified motifs appear to be the novel sequence elements that are different from the known imprinting elements. The newly discovered motifs can improve our understanding of the mechanism of genomic imprinting and their roles in human cancers and diseases.

Methods

Data source

We collected a list of 41 imprinted genes in human from the literature. The full list of the 41 imprinted genes can be found in Supplemental Table 1. The genomic DNA sequences of the imprinted genes were retrieved from NCBI's NT

sequences, which can be downloaded from <ftp://ncbi.nlm.nih.gov/genomes/>. The mouse homologous genes were determined from <ftp://ftp.ncbi.nih.gov/pub/HomoloGene/hmglg.ftp> and the literature. For each pair of human–mouse homologous genes, we collected genomic DNA of the entire gene as well as 10-kb flanking sequences. We used the program PipMaker [15] to search conserved regions for each pair of human–mouse homologous imprinted gene. Genomic sequences were masked for repeats by RepeatMasker (<http://ftp.genome.washington.edu/RM/RepeatMasker.html>) before PipMaker analysis. GENSCAN was also performed to exclude exons [21]. Segments of more than 50 bp in length and 75% nucleotide identity were considered human–mouse conserved sequences. The threshold of 50 bp and 75% homology generated optimal numbers of conserved sequences for searching motifs. These conserved sequences were then used for motif search (described in the next section).

Motif search and analysis

For a given imprinted gene, the conserved sequences were concatenated. Concatenation was carried out for each of the upstream, intron, and downstream sequences. We downloaded MEME and MAST programs from <http://meme.sdsc.edu/meme/website/meme-download.html> and

installed them on a local server. The MEME program was run to search common motifs in these conserved sequences among the imprinted genes. The MEME program produces a consensus sequence and a motif weight matrix. We used the MAST program, using the motif weight matrix from the MEME program, to search for the motif in the 24 imprinted genes as well as in 128 nonimprinted genes (data can be found in Supplemental Tables 1 and 2). Fisher's exact tests were used to test association of the motif with imprinted genes. Based on the presence or absence of each motif in the 24 imprinted genes and the 128 nonimprinted genes, we conducted Fisher's exact test based on the following table:

	Motif present	Motif absent
Imprinted	n11	n12
Nonimprinted	n21	n22

For each motif, the numbers n11 and n21 can be found in the third and fourth columns in Table 1, respectively. The numbers n12 and n22 were calculated from $(24 - n11)$ and $(128 - n21)$, respectively.

Logistic regression analysis of imprinting motifs

Based on the training set of the 24 known imprinted genes and the 128 nonimprinted genes, we generated a logistic regression model to score imprinted genes. We initially had 16 motifs as predictor variables for the model. However, when all 16 motifs were used to build a logistic regression model, the iteration process to find the coefficients of the model was not convergent. We excluded motif 4, motif 5, and the motif 15 because their p values in Fisher's exact test were greater than 0.01. We also excluded motif 11 since it was underrepresented in the imprinted genes. We started a model with 12 motifs. An input vector to the model is a feature vector for a gene indicating whether each of these 12 motifs is present or absent in the gene. The response of the model is the probability of the gene being an imprinted gene. We performed the stepwise model selection by minimizing AIC criterion and found 6 optimal motifs (motifs 3, 7, 10, 12, 13, and 16, see Table 1) as input variables for a logistic regression model to score imprinted genes. As we reduced the number of the predictor variables from 12 to 6, the AIC of the corresponding model dropped from 43.2 to 34. The AIC for the model with 5 motifs is 37 (Fig. 3D). Therefore, the 6 motifs are optimal predictors from the AIC point of view. In fact, we computed the AIC for every possible subset of the 12-motif set. The 6-motif set (3,7,10,12,13,16) has the minimum AIC. Using these 6 motifs as input variables, we estimated the model

$$p = 1 / (1 + \exp(7.1 - 4.8 \times M3 - 12.2 \times M7 - 4.2 \times M10 - 4.9 \times M12 - 12.1 \times M13 - 12 \times M16)),$$

where M3, M7, M10, M12, M13, and M16 indicate motif 3, motif 7, motif 10, motif 12, motif 13, and motif

16, respectively, and \times denotes multiplication and \exp refers to exponential function. The performance of this model on the training set is indicated by three indices: accuracy = 98%, sensitivity = 92%, and specificity = 99%.

References

- [1] R.D. Nicholls, J.H. Knoll, M.G. Butler, S. Karam, M. Lalande, Genetic imprinting suggested by maternal heterodisomy in nondeletion Prader–Willi syndrome, *Nature* 342 (1989) 281–285.
- [2] J. Clayton-Smith, M.E. Pembrey, Angelman syndrome, *J. Med. Genet.* 29 (1992) 412–415.
- [3] M. Mannens, J.M. Hoovers, E. Redeker, M. Verjaal, A.P. Feinberg, P. Little, M. Boavida, N. Coad, M. Steenman, J. Blik, et al., Parental imprinting of human chromosome region 11p15.3–pter involved in the Beckwith–Wiedemann syndrome and various human neoplasia, *Eur. J. Hum. Genet.* 2 (1994) 3–23.
- [4] S. Rainier, L.A. Johnson, C.J. Dobry, A.J. Ping, P.E. Grundy, A.P. Feinberg, Relaxation of imprinted genes in human cancer, *Nature* 362 (1993) 747–749.
- [5] O. Ogawa, M.R. Eccles, J. Szeto, L.A. McNoe, K. Yun, M.A. Maw, P.J. Smith, A.E. Reeve, Relaxation of insulin-like growth factor II gene imprinting implicated in Wilms' tumour, *Nature* 362 (1993) 749–751.
- [6] J.S. Sutcliffe, M. Nakao, S. Christian, K.H. Orstavik, N. Tommerup, D.H. Ledbetter, A.L. Beaudet, Deletions of a differentially methylated CpG island at the SNRPN gene define a putative imprinting control region, *Nat. Genet.* 8 (1994) 52–58.
- [7] A. Wutz, O.W. Smrzka, N. Schweifer, K. Schellander, E.F. Wagner, D.P. Barlow, Imprinted expression of the Igf2r gene depends on an intronic CpG island, *Nature* 389 (1997) 745–749.
- [8] M.P. Lee, M.R. DeBaun, K. Mitsuya, H.L. Galonek, S. Brandenburg, M. Oshimura, A.P. Feinberg, Loss of imprinting of a paternally expressed transcript, with antisense orientation to KVLQT1, occurs frequently in Beckwith–Wiedemann syndrome and is independent of insulin-like growth factor II imprinting, *Proc. Natl. Acad. Sci. USA* 96 (1999) 5203–5208.
- [9] A.P. Feinberg, Imprinting of a genomic domain of 11p15 and loss of imprinting in cancer: an introduction, *Cancer Res.* 59 (1999) 1743s–1746s.
- [10] M.P. Lee, S. Brandenburg, G.M. Landes, M. Adams, G. Miller, A.P. Feinberg, Two novel genes in the center of the 11p15 imprinted domain escape genomic imprinting, *Hum. Mol. Genet.* 8 (1999) 683–690.
- [11] W.W. Wasserman, M. Palumbo, W. Thompson, J.W. Fickett, C.E. Lawrence, Human–mouse genome comparisons to locate regulatory sites, *Nat. Genet.* 26 (2000) 225–228.
- [12] P. Onyango, W. Miller, J. Lehoczy, C.T. Leung, B. Birren, S. Whealan, K. Dewar, A.P. Feinberg, Sequence and comparative analysis of the mouse 1-megabase region orthologous to the human 11p15 imprinted domain, *Genome Res.* 10 (2000) 1697–1710.
- [13] M. Paulsen, S. Takada, N.A. Youngson, M. Benchaib, C. Charlier, K. Segers, M. Georges, A.C. Ferguson-Smith, Comparative sequence analysis of the imprinted Dlk1–Gtl2 locus in three mammalian species reveals highly conserved genomic elements and refines comparison with the Igf2–H19 region, *Genome Res.* 11 (2001) 2085–2094.
- [14] C. Charlier, K. Segers, D. Wagenaar, L. Karim, S. Berghmans, O. Jaillon, T. Shay, J. Weissenbach, N. Cockett, G. Gyapay, M. Georges, Human–ovine comparative sequencing of a 250-kb imprinted domain encompassing the callipyge (clpg) locus and identification of six imprinted transcripts: DLK1, DAT, GTL2, PEG11, antiPEG11, and MEG8, *Genome Res.* 11 (2001) 850–862.
- [15] S. Schwartz, Z. Zhang, K.A. Frazer, A. Smit, C. Riemer, J. Bouck, R.

- Gibbs, R. Hardison, W. Miller, PipMaker—a Web server for aligning two genomic DNA sequences, *Genome Res.* 10 (2000) 577–586.
- [16] T.L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2 (1994) 28–36.
- [17] T.L. Bailey, M. Gribskov, Combining evidence using p-values: application to sequence homology searches, *Bioinformatics* 14 (1998) 48–54.
- [18] S. Lo, Z. Wang, Y. Hu, H.H. Yang, S. Gere, K.H. Buetow, M.P. Lee, Allelic variation in gene expression is common in the human genome, *Genome Res.* 13 (2003) 1855–1862.
- [19] J.M. Greally, Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome, *Proc. Natl. Acad. Sci. USA* 99 (2002) 327–332.
- [20] M. Paulsen, A.C. Ferguson-Smith, DNA methylation in genomic imprinting, development, and disease, *J. Pathol.* 195 (2001) 97–110.
- [21] C. Burge, S. Karlin, Prediction of complete gene structures in human genomic DNA, *J. Mol. Biol.* 268 (1997) 78–94.